Response to RC1 comments on "A new skill score for ensemble flood maps: assessing spatial spread-skill with remote sensing observations".

Many thanks to reviewer 1 for your useful feedback and time reviewing our paper. Our responses to numbered comments in black are detailed below in *blue italics.* We have listed the original line numbers from the reviewers' minor comments. We include a marked-up copy of the revised manuscript where text removed is in strikeout font and additions are underlined.

Response to RC1 comments

The following may help the reader understand the methods:
1) referring to figures and metrics consistently, and frequently referencing back to the source equations (e.g., 'this figure shows the agreement scale calculated with Eq. 7').
Referencing of the source equations have been added to the caption of Fig. 8 and Sections 2.1, 2.2 and 2.3 following comments in the pdf supplement.
2) providing an intuitive explanation for the metrics (e.g., 'high agreement scales means the two grids agree on the proximity of flooding in this region').
Added to Section 2.2 'A small value for the agreement scale means that the two arrays being compared are very similar (spatially) at a specific location, whereas a large value for the agreement scale means that the two arrays being compared are dissimilar.'
3) consistent notation (i.e., use upper case for grids and lower case for variables, use brackets for operators, define all variables and grids). *Amendments have been made to the variables used to improve consistency following comments in the pdf supplement. All data arrays are now labelled using upper case variables and we have defined S and α.*
4) a workflow diagram for computing each metric (along with a toy example). *In our previous paper (Hooker et al., 2022, which we direct readers to in Sections 2.1 and 2.2) we presented a simple "toy" example for the FSS along with a case study where agreement scale maps were plotted. This research article is presented as a first example of applying a spatial spread-skill metric to forecast flood maps and as such the whole article provides the detailed methods and an application to aid understanding.*

5) Similarly, providing some simpler alternative metric for skill, then demonstrating how the proposed sophisticated method adds some useful information would greatly improve the manuscript. Without this, it reads like an overly complicated method has been imported from meteorology when something much simpler and easier to understand would do (e.g., critical success index of the median grid). I mentioned this in my original comments, and the authors replied that their method is too novel. However, in the conclusion they mention other methods for 'presentation' which could easily be extended for (simple) validation
Added to introduction 'In a recent study, a scale-selective approach was developed and applied to evaluate a deterministic flood map forecast where comparisons were made against conventional binary performance measures (Hooker et al., 2022a). 'A scale-selective approach to flood map evaluation was found to have several benefits over conventional binary performance measures. These include overcoming the

double penalty impact problem when validating at higher spatial resolutions and accounting for the impact of the flood magnitude on the skill score.'

6) My comment about evaluating the accuracy of the SAR layer has not been addressed. The authors now provide useful information on the SAR layer, but no evaluation is provided. What if this SAR layer is from the wrong day? or is being confused by vegetation? How would this effect the case study? I understand there is no better data available, but this is not justification to assume the SAR layer is accurate. If the authors want to stick with this case study (rather than switch to one with more reliable validation data), the text needs to be revised to acknowledged that the validation is being performed against dubious data. For example, the discussion on line 410 needs to consider this.

The goal of our manuscript is to establish a new method for spatial evaluation of ensemble flood forecasts. The method and its uses are the main point of the paper, not the performance of the specific flood forecasting system that we use as an example. Hence, in our opinion our choice of case study is a good one as it allows us to demonstrate the main features of the method. Of course, any users of our new method should take into account the uncertainties of the verifying observations when interpreting their results. Indeed, we do discuss the uncertainties in the verifying observations that we use. The limitations of SAR data are discussed in detail in the introduction and their use as flooding extent observations is well established in research and operational flood risk management (e.g. Grimaldi et al. 2016, added as an additional reference). In Section 3.3 we state the time and date of the SAR acquisition, and this can be found in the data repository. We also mention that vegetation and urban areas have been flood filled using morphological closing to allow a fairer comparison.

Added to Section 3.3 'Additionally, a flood mask, indicating areas where flood detection using SAR data is not currently possible (at the Sentinel-1 spatial resolution) could be used to exclude areas from the evaluation process (note that this was not possible for this case study, since this information was not available in 2017).'

7) If I understand the Flood Foresight pipeline, the return period maps (coming from Glofas?) are what is driving heterogeneity in the inundation ensemble. Obviously, it's not so interesting to apply your metric to this, but the evaluation would be more transparent (and the nature of the ensemble more clear) if these maps were also provided (or at least described in some way).

The introduction (first paragraph) explains that the ensemble numerical weather prediction model is driving heterogeneity in the inundation ensemble. The flood map library contains maps calculated for a particular return period (or through interpolation). In practice, the forecast/observed return period is not the same at all sub-catchments so it would not be meaningful to evaluate the library flood maps against the SAR-derived flood map.

8) There are some fixed values discussed in the methods ('binary threshold', 'Slim') but I was not able to find these discussed (or specified) for the case study.

Thank you, a good spot, the value of Slim for the case study has been added.

Response to RC1 comments in supplementary attachment

Changes and additions in response to most of these minor comments can be found in the marked version of the revised manuscript. A few cases where we have not made changes are detailed below with original line numbers listed.

1) L3 'Insurers… not a typical consumer of flood forecasts…' *Flood inundation maps and flood forecasts are commonly and increasingly used by insurance companies (e.g. https://redcross.eu/projects/forecast-based-financing).*
2) L25 '…weather forecast… I thought we were talking about fluvial flooding. *The previous sentence links fluvial flooding to rainfall uncertainties and atmospheric initial condition uncertainty.*
3) L36 'predictability' *We maintain that this is the correct terminology and is used in multiple previously published papers (e.g. Dey et al., 2016), as well as the title of ECMWF training courses e.g., https://www.ecmwf.int/en/learning/training/predictability-and-ensemble-forecasting.*
4) L121. 'does it have to be compared to a remote sensing product? can't it be compared to a hydrodynamic product?' *We maintain that the forecast should be validated against an independent observation. Note that the observed river streamflow is unknown.*
5) L306 'If I understand, these return period maps are really what is driving heterogeneity in the ensemble. It would be helpful to see these return period maps. Are they homogeneous? Or is there a lot of variance? I suggest adding this as a supplement.' *The main driver of the heterogeneity is the flood map selection for each sub-catchment. This is clearer now with the addition of the new Figure 4 (see RC2 comment 2).*
6) L332. 'how did you use the uncertainty?' *The work (and flood case study) for this manuscript took place before the GFM SAR uncertainty information became available (see response to main comment 6).*
7) Fig 4. 'Can you sort these from 'most' to 'least' flooded?' *We present these as typical ensemble forecast products as an example of the difficulty in analysing the outputs. As such we prefer to keep them in the current format.*

Response to RC3 comments

Many thanks to reviewer 3 for your useful feedback and time reviewing our paper.

General comments

1. The library-approach used to generate the forecast flood maps is not well described which makes it difficult to assess the validity of the case study.

a. For instance, from the authors explanation of the flood mapping framework it seems like JFlow and RFlow hydrodynamic models are both used to produce flood maps for different return periods but it's not clear how they are used. Is it that Jflow is used for some catchments and RFlow is used for others? Are they coupled (i.e. the output of one

of the models is used as input for the other)? The authors should provide more details on how flood maps are produced for the study area.

We have made it clearer in the manuscript that JFlow is used where a DTM is available and RFlow is used where a DTM is unavailable.

Also added 'Flood maps were pre-computed for the domain of interest (Fig. 2) using a DSM and RFlow.'

b. In line 285, the authors mention that streamflow forecasts and observations are used to select the "most appropriate map" for each location but no other details are provided regarding this process. Is the most appropriate map selected based on discharge/stage observations and forecast at gauging stations? What gauging stations are used to select the most appropriate maps? If there are multiple gauging stations in the study area, how are the most appropriate maps selected?

For this international application of Flood Foresight, only forecast streamflow is used to select the flood maps. Removed 'river gauge data (both historical and real-time)'.

c. In line 305 the authors mention that: "Each of the GloFAS grid cells are linked to sub-catchments in the Flood Foresight system". What sub-catchments are the authors referring to? How is the study area divided into different sub-catchments?

Added 'Flood foresight is set up for a region by dividing the river basin into sub-catchments using the HydroBASINS data-set (level 12) (Lehner et al., 2014). Flood Foresight takes gridded inputs of ensemble forecast streamflow and uses these to select the most appropriate flood map for each sub-catchment. These are mosaicked together and forecasts of ensemble flood maps are produced daily, out to ten days ahead.'

d. In line 306, the authors mention that "The simulation library flood maps are selected when the forecast streamflow exceeds a return period threshold level". At which points in the study area is this condition evaluated?

Following a reference to the new Figure 4 (see response to comment 2 below) 'GloFAS outputs a gridded (approximately 10 km spatial resolution) ensemble forecast of river streamflow (Fig. 4). Each of the GloFAS grid cells are linked to the sub-catchments in the Flood Foresight system.'

We add 'The simulation library flood maps are selected when the forecast streamflow exceeds a return period threshold level within each sub-catchment. The RP threshold levels are calculated using ERA5 reanalysis data (Harrigan et al., 2020).'

2. In addition to the concerns about the explanation of the flood mapping approach, I have a concern with the way the hydrologic/hydraulic modeling is presented. It's hard to assess the results in the study case without a figure showing the hydrologic characteristics of the study are. I would like to see a figure with: basin and sub-basins, the river network, terrain elevation, gauging stations, streamflow forecast points, and all other details necessary to understand how the study area is modeled.

We have added an additional figure (new Fig. 4) showing the GloFAS grid cells (forecast streamflow points), the sub-catchments and the permanent water bodies. The closest gauging station is located downstream of the DOI and is marked on Figure 2. We reference the DSM used to create the flood maps in Section 3.2. Unfortunately, the commercial data license for the DSM does not permit us to include this on the new figure.

3. The interpolation of flood maps between return periods might be affecting the spatial skill analysis results. If I understand correctly in the Flood Foresight framework (Mason et al., 2021), between two simulated flood maps, the water depth at each flooded cell is

interpolated and 5 different flood maps are generated where water depth at flooded cells linearly changes. Thus, for interpolated maps, water depth changes but flooded cells stay the same as in the simulated flood map. This implies that, for example, the 50-, 60-, 70-, 80-, and 90-year return period flood maps contain the same number and location of flooded cells. Then, with this flood mapping approach a spatial skill analysis based on neighborhood might not be appropriate.
The interpolation process reflects both the change in depth and extent from one modelled return period map to next, therefore each interpolated map can have a different number of flooded pixels. Added '…of both flood depth and extent'.
4. It's not clear how the categorical scale map (Figure 7) is constructed. How is over- and under-prediction obtained from the scale agreement metric alone? Please give more details. Also, according to equation 7, agreement scale is always positive. Why are there negative values in Figure 7?
Added 'In the contingency table under-predicted cells are set to +1, over-predicted cells are set to -1, correctly predicted flooded cells are assigned NaN and correctly predicted unflooded cells are set to 0…(by element-wise array product).'
5. From visual inspection of Figure 4, ensemble member 1 which is the best performing member according to the FSS analysis, has a tendency to over-predict floods. Similarly, the combined ensemble (ensall) is among the best performing forecasts according to FSS. I wonder if forecasted flood maps that over-predict floods yield a better FSS vs n relationship. Is this the case? What other members are in the top cluster in Figure 5?
There is no evidence in the literature to suggest that the FSS score is biased towards over-prediction. Both the ens_{all} and ens_{median} exceed FSS_T at a similar score and the worst performing ensemble member (ens_{21}) has a large area of over-prediction (similar to ens_{all}) but is missing the flooding observed to the north (area 1 Fig. 9, discussed in Section 4.3.) that is captured by ens_{all}.

Specific comments

1. Title: Is this really a "new" skill score? The authors already presented the spatial forecast skill score (FSS) in a previous publication (Hooker et al. 2022).
This paper presents ensemble skill scores whereas our previous work (Hooker et al 2022) was for a single "deterministic" forecast. Changed title to 'Assessing the spatial spread-skill of ensemble flood maps with remote sensing observations'.

2. Lines 133: The use of word "depending" is not correct in this sentence. Change "model" by "modelled"
Changed to 'A potential maximum $MSE_{n(ref)}$ depends on the fraction of flooding in the domain for the modelled and observed fields and is calculated as:'
3. Line 140: Explain in more detail how FSS_T formula is obtained. This threshold on FSS is important as results presented in section 4.1 all depend on this threshold. A complete explanation and strong justification for FSS_T formula should be provided.
Added 'A recent study by Skok et al., 2018 investigated the sensitivity of the calculated skilful scale to the constant value (0.5) in Eq. (4), and found that 0.5 gave meaningful results compared with the measured displacement.'
4. Line 180: Improve explanation of how categorical scale maps are generated.

Please see response to main comment 4.

5. Figure 3: Was this figure prepared by the authors or by JBA consulting? If the figure was prepared by JBA consulting, credit should be given in the figure caption. If the figure was prepared by the authors there's no need to include Flood Foresight or JBA logos.

Added 'Prepared by JBA Consulting.' We note that two of the authors work for JBA Consulting.

6. Line 323: Please reword sentence "So that…".

Changed to 'In order to evaluate the flood prediction accuracy alone…'

7. Line 395: The sentence discussing Fig. 9 is confusing here. I recommend moving this sentence after presenting results for Fig. 8.

We prefer to introduce the figures here, as the next paragraph discusses both figures together. Changed reference to '…Spatial Spread-Skill (SSS) map (derived from Fig. 9, presented in Fig. 10)'.